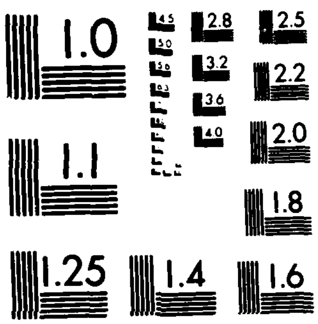
AD-A127 705 THE IMPORTANCE OF PRACTICE IN THE DEVELOPMENT OF STATISTICS(U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER G E BOX JAN 83 MRC-TSR-2471 1/1

UNCLASSIFIED DAAG29-80-C-0041 F/G 12/1 NL

END
DATE
FILMED
5-83
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD A127705

MRC Technical Summary Report #2471

THE IMPORTANCE OF PRACTICE IN THE DEVELOPMENT OF STATISTICS

George E. P. Box

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

January 1983

(Received November 9, 1982)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

THE IMPORTANCE OF PRACTICE IN THE DEVELOPMENT OF STATISTICS*

George E. P. Box

Technical Summary Report #2471
January 1983

ABSTRACT

The paper shows how application and consideration of the scientific context in which Statistics is used can initiate important advances such as: least squares, ratio estimators, correlation, contingency tables, studentization, experimental design, the analysis of variance, randomisation, fractional replication, variance component analysis, bioassay, limits for a ratio, quality control, sampling inspection, non-parametric tests, transformation theory, ARIMA time series models, sequential tests, cumulative sum charts, data analysis plotting techniques, and a resolution of the Bayes - frequentist controversy.

It appears that advances of this kind are frequently made because practical context reveals a novel formulation which eliminates an unnecessarily limiting framework.

AMS (MOS) Subject Classifications: 62-03, 01A99, 62A15, 62A20

Key Words: Practice, theory, least squares, ratio estimators, correlation, contingency tables, studentization, experimental design, the analysis of variance, randomisation, fractional replication, variance component analysis, bioassay, limits for a ratio, quality control, sampling inspection, non-parametric tests, transformation theory, ARIMA time series models, sequential tests, cumulative sum charts, data analysis plotting techniques, Bayesian inference, sampling theory.

Work Unit Number 4 (Statistics and Probability)

*A lecture especially prepared for, and videotaped by, the American Statistical Association in their program for filming distinguished statisticians.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

Statistics is concerned with the analysis and generation of scientific data. As might be expected, therefore, much valuable research in Statistics has been initially motivated by practical problems emerging from their scientific context. Examples are given from the work of Gauss, Laplace, Galton, Karl Pearson, Gosset, Fisher, Yates, Youden, Finney, Plackett and Burman, Tippett, Daniels, Egon Pearson, Shewhart, Dodge, Tukey, Bartlett, Wilcoxon, Yule, Holt, Winters, Wald, Barnard, Page and Cuthbert Daniel.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

THE IMPORTANCE OF PRACTICE IN THE DEVELOPMENT OF STATISTICS*

George E. P. Box

1. INTRODUCTION

The importance of practice in guiding the development of Statistics hardly needs emphasis. And yet I think it is worth examination. For statistical methods and statistical theory, like so many other things, evolve by a process of natural selection. Least squares, invented at the beginning of the 19'th century, is alive and well but the coefficient of colligation is now seldom used. For development to occur both appropriate tools and motivation are needed. The tools are mathematics, numerical analysis and computation. An important motivation is the practical need to solve problems. Tools and motivation interact of course. For example the existence of fast computers is encouraging the development of new statistical methods which would have been quite impossible without them, and which presage further theoretical development. Again, advance must sometimes wait on knowledge of appropriate mathematics. Thus Fisher's ability to solve the distributional problems of correlation and of the linear model rested strongly on his facility with n-dimensional geometry which his contemporaries lacked.

It would be hard to argue, however, that any one deficiency in the tool-kit is disastrous. Thus least squares, although, according to Gauss, fully known to him in 1796, could require calculations which were dauntingly burdensome until the onset of modern computers in the 1950's. Again, Galton, Gosset, and Wilcoxon, pioneers respectively in the concepts of correlation, studentization and non-parametric tests did not regard themselves as

*A lecture especially prepared for, and videotaped by, the American Statistical Association in their program for filming distinguished statisticians.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

particularly competent mathematicians. In particular, Gosset's derivation of the sampling distribution of what we now call the t-statistic must surely stand as the nadir of rigorous argument. But he did get the right answer; and he was first.

My theme then will be to illustrate how practical need often leads theoretical development. Early examples would be the development of the probability calculus which was closely bound up with the desirability of winning at games of chance, the introduction of least squares by Gauss to reconcile astronomical and survey triangulation measurements, and the invention by Laplace of ratio estimators to determine the population of France (at the request of Napoleon)(1).

Let us consider some of the children of necessity produced in more modern times.

2. FURTHER EXAMPLES OF THE PRACTICE-THEORY INTERACTION

In the middle of the nineteenth century the impact of Darwin's ideas was dramatic. But Darwin, although an intellectual giant, had little mathematical ability. To Francis Galton the challenge was obvious; the rightness and further consequences of Darwin's ideas must be demonstrable using numbers. For example, given that offspring varied about some kind of parental mean why, with each new branching of a family tree, did not variation of species continually increase? The answer to this practical question lay(2), he discerned, in the regression towards the mean implied by the bivariate normal surface which ensures that, on the average, sons of six feet fathers are less than six feet. Again, it was the need which he perceived of a measure of the intensity of the partial similarities between pairs of relatives that led to his introducing the concept of correlation, an idea taken up with great enthusiasm and further developed by Karl Pearson.

Pearson was a man of enormous energy and very wide interest including social reform and the general improvement of the human condition. He was, however, conscious of the fact that, in deciding what kind of reforms ought to be sought, good intentions, although necessary, were not always sufficient. A course of action based on the accepted belief that alcoholism in parents produced mental deficiency in children might be ill advised if, as he demonstrated[3], that belief was contradicted by the data. Obviously correlation might be useful in such studies, but other measures and tests of association were needed for qualitative variables. Pearson developed such tools, and in particular his χ^2 test for contingency tables.

Karl Pearson's methods were developed mainly for large samples and as they stood they did not meet the practical needs of W. S. Gosset when he came to study statistics for a year at University College, London, in 1906. Gosset had graduated from Oxford in Chemistry and had gone to work for Guinnesses, following their policy begun in 1893 of recruiting scientists as brewers. He soon found himself faced with the analysis of small sets of observations coming from the laboratory, from field trials, and from the experimental brewery of which he was placed in charge in 1905[4].

The general problem Gosset faced was how to deal with unknown nuisance parameters, and specifically the unknown standard deviation in the comparison of means. The method then in use was to substitute some sort of estimate for an unknown nuisance parameter and then to assume that one could treat the result as if the true value had been substituted. While this might provide an adequate approximation for large samples it was clearly inadequate when the sample was small. Furthermore he did not find nor expect to find that there was much interest in his problems. (He wrote to Fisher of the t-tables, "you are probably the only man who will ever use them.")[5] It must have been clear to him that, at that time, if anyone was to do anything about small samples it would have to be himself.

Gosset's invention of the t test was a milestone in the development of statistics because it showed how, by studentization, account might be taken of the uncertainty in an estimated nuisance parameter. It thus paved the way for an enormous expansion of the usefulness of statistics, which could now begin to provide answers for agriculture, chemistry, biology and many other subjets where small samples, rather than large samples, were the rule.

Fisher, as he always acknowledged, owed a great debt to Gosset, both for providing the initial clue as to how the general problem of small samples might be approached, and also for mooting the idea of statistically designed experiments.

When Fisher went to Rothamsted in 1919 he was one of a number of young scientists newly recruited by Russell. He was immediately confronted with a massive set of data - rainfall every day, and harvested yields every year, for 13 Broadbalk plots that had been fertilized in the same pattern for over 60 years. As might be expected his analyses were not routine[6][7]; he introduced distributed lag models, orthogonal polynomials, an early form of the analysis of variance, and the distribution of the multiple correlation coefficient. Also to check the fit of his model he considered the properties of residuals. Furthermore he devised ingenious methods for lightening the burdensome calculations which had to be made on a desk calculator. But the most important outcome of this "raking over the muck-heap" as he called it, and of analyzing other field experiments which he had had no part in planning, came from the very deficiencies which these data presented. It was the invention of experimental design.

How, he was soon led to ask, might experiments be conducted so that they unequivocally answered the questions posed by the investigator. One can clearly see the ideas of randomisation, replication, orthogonal arrangement, blocking, factorial designs, measurement of interactions, confounding, all developing in response to the practical necessities of field experimentation[8].

Design and analysis came to play complementary roles in Fisher's thinking, so that over the period 1916-1930 we can see the Analysis of Variance first hinted at, and then developed and adapted to accompany the analysis of each new design. It is in 1923 that, in a paper with Miss Mackenzie[9], the analysis of variance first appears in the tabular form with which we are all familiar. But the object of the investigation was to solve an agricultural problem and it is typical of Fisher that there is no reference in the title of the paper to the analysis of variance or to the other new statistical ideas it contains. The paper is called, "The manurial response of different potato varieties." In it we are introduced, not only to the analysis of variance for a replicated two-way table, but also to its partial justification by randomisation theory, rather than Normal theory. In addition he presents a method of analysis (only recently rediscovered)[10] using models which are nonlinear in the parameters.

There now existed at Rothamsted a center where careful statistical planning was going into the process of generation and analysis of data coming from a host of important problems. Fisher left Rothamsted in 1933 and was succeeded by Yates who had come two years earlier and was not only a mathematician but had had much practical experience in least squares calculations in geodetic survey work. Yates[11] made many important advances. In particular he further developed factorial designs and confounding, invented many new designs, including balanced incomplete block arrangements, and showed how to cope, when, as sometimes happened, things went wrong and there were missing data.

These ideas found wide application and inspired much new research. For example, Jack Youden[12], then working at the Boyce Thomson Institute, was involved in an investigation of the infective power of crystalline preparations of the tobacco-mosaic virus. Not only did the test plants vary from one to the other in their tendency to infection but leaves from the same plant varied depending on their position and each plant could not be relied upon to provide more than five experimental leaves. In response Youden invented what came to be called the Youden Square, a design which stands in the same relationship to the latin square as the balanced incomplete block does to the randomised block design.

Later, another important development coming from Rothamsted was fractional replication. Fisher had pointed out[13] that in suitable circumstances, adequate estimates of error could be obtained in large unreplicated factorials from estimates of high order interactions which might be assumed negligible. In 1945 David Finney[14], responding to the frequent practical need to maximise the number of factors studied per experimental run, further exploited this possible redundancy by introducing fractional factorial designs. These together with another broad class of orthogonal designs developed independently by Robin Plackett and Peter Burman[15] in response to war time problems, have since proved of great value in industrial experimentation. An isolated example of how such a design could be used for screening out a source of trouble in a spinning machine had been described as early as 1934 by L.H.C. Tippett of the British Cotton Industry Research Association[16]. The arrangement was a 125'th fraction of a 5^5 design!

It seems that wherever a good source of problems existed in the presence of a suitably agile mind new developments were bound to occur. Thus the pressing problem of drug standardisation in the hands of J. H. Gaddum[17], C. I. Bliss[18] and (again) D. J. Finney[19], gave rise to modern methods of bioassay using probits, logits and the like. While in 1940 a study of the standardisation of insulin led Edgar Fieller[20] working for Boots Pure Drug Company to a resolution of the problem of finding confidence limits for a ratio and for the solution of an equation whose coefficients were subject to error.

Earlier, Henry Daniels[21], then a statistician at the Wool Industries Research Association, showed how variance component models could be used to expose those parts of a production process responsible for large variations. Variance component analysis has since proved of enormous value in the process industries and elsewhere.

Daniels' contribution was one in a series of papers on industrial statistics read in the 1930's before what was then called the Industrial and Agricultural Research Section of the Royal Statistical Society. A leading spirit in getting this section moving was Egon Pearson whose ideas greatly influenced, and were influenced by, this body. In particular he liked data analysis and graphical illustration and used it effectively[22] to illustrate Daniels' conclusions.

An important influence on Egon Pearson was the work of Walter Shewhart[23] on quality control. This work and that on sampling inspection by Harold Dodge[24] heralded more than half a century of statistical innovation coming from the Bell Telephone Laboratories. This has led most recently to rekindling of interest in data analysis in a much needed revolution led by John Tukey[25][26].

Another innovator guided by practical matters was Frank Wilcoxon, a statistician for the Lederle Labs of the American Cyanamid Company. Just after the second world war in the age of desk calculators, Frank found himself confronted by the need to make thousands of tests on samples from the pharmaceutical research then in progress. He said it was the need for quickness rather than anything else which led to the famous Wilcoxon tests[27], the precursors of much subsequent research on non-parametric methods.

M. S. Bartlett's contributions to statistics are legion, but one would certainly suppose that his early contributions to the theory of transformation[28] had much to do with the fact that when he was statistician at the Jealotts Hill agricultural research station of Imperial Chemical Industries, he was much concerned with the testing of pesticides and so with data that appeared as frequencies or proportions.

Another clear example of the practice-theory interaction is seen in the development of parametric time series models. In 1927 Udny Yule was trying to understand what was wrong with William Beveridge's analysis of wheat price data. The fitting of sine waves of different frequencies by least squares had revealed significant oscillations at strange and inexplicable periods. Yule suggested that such series ought to be represented, not by a deterministic function subject to error, but by a dynamic system (represented by a linear difference equation)[29] responding to a series of random shocks - this model was likened to a pendulum being periodically hit by peas from a pea shooter. Yule's revolutionary idea, with important further input from Slutsky[30], Wold[31] and others, was the origin of autoregressive-moving average models.

Unfortunately the practical use of these models was for some time hampered by an excessive emphasis on stationary processes which vary in equilibrium about a fixed mean. The requirement for stationarity is that the characteristic polynomial for the autoregressive part of the model must have all its zeroes outside the unit circle. Many of the series arising in business and economics did not, however, behave like realisations from such a stationary model. Consequently, for lack of anything better, operations research workers led by Holt[32] and Winters[33] began in the 1950's to use the exponentially weighted moving average of past data and its extensions for forecasting series of this kind. This weighted average was introduced at first on purely empirical grounds - it seemed sensible to monotonically discount the past and it seemed to work reasonably well. However, in 1960, Muth[34] showed, rather unexpectedly, that this empirically derived statistic was an optimal forecast for a special kind of autoregressive - moving average model. This model was not stationary. Its autoregressive polynomial had a root on the unit circle. The general class of models with roots on the unit circle, where stationarity would forbid them, later turned out to be extremely valuable for representing many kinds of practically occurring series, including seasonal series.

The second world war was of course a stimulus to all kinds of invention. Allen Wallis[35] has described the dramatic consequence of a practical query made by a serving officer about a sampling inspection scheme. The question was of the kind "Suppose, from a sample of twenty items, three is the critical number of duds, if it should happen that the first three components tested are all duds, why do we need to test the remaining seventeen?" Wallis and Milton Friedman were quick to see the apparent implication of this question, that "super-powerful" tests were possible!

However, their suggestion that Abraham Wald be invited to work on the problem was resisted for some time. It was argued that this would clearly be a waste of Wald's time, because to do better than a most powerful test was impossible. What the objector had failed to see was that the test considered was most powerful only if it was assumed that n was fixed, and what the officer had seen was that n did not need to be fixed. It is well known how this led to the development of sequential tests[36]. It is heartening that this particular happening even withstood the scientific test of repeatability, for at about the same time and with similar practical inspiration, sequential tests (of a somewhat different kind) were discovered independently in Great Britain by George Barnard[37].

Nor was this the end of the story. Some years later Ewan Page, then a student of Frank Anscombe, while considering the problem of finding more efficient quality control charts, was led to the idea of plotting the cumulative sum[38] of deviations from the target value.

The concept was further developed by Barnard in 1959[39] who introduced the idea of a V mask to decide when action should be taken. The procedure is identical to a backwards running two-sided sequential test. Cusum charts have since proved to be of great value in the textile and other industries. In addition, this graphical test has proved its worth in the "post mortem" examination of data where it can point to the dates on which critical events may have occurred. This sometimes leads to discovery of the reason for the events.

A pioneer of graphical techniques of a different kind is Cuthbert Daniel, an industrial consultant who has used his wide experience to make many contributions to statistics. An early user of unreplicated and fractionally replicated designs, he was concerned with the practical difficulty of estimating error. In particular he was quick to realize that higher order interactions sometimes do occur and when they do it is important to isolate and study them. His introduction of graphical analysis of factorials by plotting effects and residuals on probability paper[40] has had major consequences. It has encouraged the development of many other graphical aids and together with the work of John Tukey has contributed to the growing understanding that at the hypothesis generation or model-modification stage of the cycle of discovery, it is the imagination that needs to be stimulated and that this can often best be done by graphical methods.

3. SOME INTERIM CONCLUSIONS

Obviously I could go on with other examples but at this point I should like to draw some interim conclusions.

I think it is possible to see important ingredients leading to statistical advance. They are

(a) the presence of an original mind that can perceive and formulate a new problem and move to its solution

(b) a challenging and active environment for that mind, conducive to discovery.

Gosset at Guinnesses; Fisher, Yates and Finney at Rothamsted; Tippett at the Cotton Research Institute; Youden at the Boyce Thomson Institute (with which organisation Wilcoxon and Bliss were also at one time associated); Daniels and David Cox at the Wool Industries Research Association; Shewhart, Dodge, Tukey and Colin Mallows at Bell Labs; Wilcoxon at American Cyanamid; Cuthbert Daniel in his consulting practice; these are all examples of such fortunate conjunctions.

Further recent examples are Don Rubin's work at E.T.S.; Jerry Friedman's computer intensive methods developed at the linear accellerator; George Tiao's involvement with environmental problems; Brad Efron's interaction with Stanford Medical School; the late Gwilym Jenkins' applications of time series analysis in systems applications; John Nelder's development of statistical computing at Rothamsted.

One message seems clear: a statistican who believes himself capable of genuinely original research can find fulfillment in a stimulating investigational environment.

Also I think it possible to understand something of the specific nature of the contribution coming from applications - frequently it is the establishment of a new frame of reference for a problem. This may involve extension, modification or even abandonment of a previous formulation. It has to be understood that statistical problems are frequently not like, for example, chess problems which may require "White to mate in three moves", given a particular configuration of the pieces. Here a solution based on the pretence that a knight can move like a queen would be unacceptable. Yet the changes in the rules that have sometimes been adopted in reformulation of statistical problems must, at the time of their introduction, have been thought of as little short of cheating. Some examples would be:

Fisher's replacement of the method of moments by maximum likelihood.

Yates' use of designs in which the number of treatments exceeded the block size.

Yule's introduction of stochastic difference equations replacing deterministic models.

Wald's and Barnard's introduction of sequential tests to replace fixed sample tests.

Page's and Barnard's introduction of quality control charts in which the cumulative sum of the deviations rather than the deviations themselves was plotted.

Finney's use of fractional, rather than full, factorials.

Fisher's use of the randomisation test to justify normal theory tests as approximations.

Daniel's and Tukey's initiation of informal graphical techniques rather than more formal procedures in data analysis.

4. A POSSIBLE RESOLUTION OF THE BAYES CONTROVERSY

One further matter that I think is greatly clarified by the practical context of its application concerns the problem of statistical inference. Here the consideration of scientific context provides, I believe, a resolution of what is sometimes called the Bayesian controversy. At its most extreme this controversy is a dispute between those who think that all statistical inferences should be made using a Bayesian posterior distribution and others who believe that sampling theory (that is, frequentist theory) has universal inferential applicability.

I have recently argued[41][42] that the Bayes-Sampling theory controversy arises because of an erroneous tacit assumption that there is only one kind of scientific inference, for which there are two candidates, whereas I believe a study of the process of scientific investigation itself shows that it requires two quite distinct kinds of scientific inference for each of which, one, and not the other, of the Bayes-Sampling candidates is appropriate. One kind of inference which may be called criticism involves the contrasting of what might be expected if the assumptions A of some tentative model of interest were true with the data y_d that actually occur. This is conveniently symbolised by subtraction: $y_d - A$. The other kind of inference, which may be called estimation, involves the combination of observed data y_d with the assumptions A of some model which is tentatively assumed true. This process is conveniently symbolised by addition: $y_d + A$.

In a statistical context, analysis of residuals, tests of fit and diagnostic checks both graphical and numerical, formal and informal, are all examples of techniques of model criticism intended to stimulate the scientist to model building and model modification, or to the generation of more relevant data should this prove desirable. These techniques must, I believe, ultimately appeal for formal justification to sampling theory.

By contrast Least Squares estimation, likelihood estimation, shrinkage estimation, robust estimation, ridge estimation, are all solutions to estimation problems which would I think be better motivated and justified by employing an appropriate model and applying Bayes theorem.

There seem to be three distinct considerations which support this dualistic view of inference: these are (a) the nature of scientific method, (b) the physiology of the brain and (c) the mathematics of Bayes theorem.

I consider them in turn.

(a) The Nature of Scientific Method:

It has for long been recognized that the process of learning is a motivated iteration between thery and practice. By practice I mean reality in the form of data or facts. In this iteration deduction and induction are employed in alternation and progress is evidenced by a developing model which by appropriate exposure to reality continually evolves until some currently satisfactory level of understanding is reached. At any given stage the current model helps us to appreciate not only what we know but what else it may yet be important to find out. It thus motivates the collection of new data appropriate to illuminate dark but possibly interesting corners of present knowledge.

We can find illustration of these matters in everyday experience, or in the evolution of the plot of any good mystery novel, as well as in any reasonably honest account of the events leading to scientific discovery.

Experimental science accelerates the learning process by isolating its essence: potentially informative experiences are deliberately staged and made to occur in the presence of a trained investigator.

The instrument of all learning is the brain - an incredibly complex structure, the working of which we have only recently begun to understand. One thing that is clear is the importance to the brain of models, where past experience is accumulated. At any given stage of experience some of the models $M_1, M_2, \ldots M_i, \ldots,$ are well established, others less so, while still others are in the very early stages of creation. When some new fact or body of facts χ_d comes to our attention, the mind tries to associate the new experience with an established model. When, as is usual, it succeeds in doing so, this new knowledge is incorporated in the appropriate model and can set in train appropriate action.

Obviously, to avoid chaos the brain must be good at allocating data to an appropriate model and at initiating the construction of a new model if this should prove to be necessary. To conduct such business the mind must be concerned with the two kinds of inferences which were mentioned previously. Namely (a) the contrasting of new facts χ_d with the assumptions A of a possible model M in an operation of criticism so stimulating induction and characterized by the subtraction $\chi_d - A$ and (b) the incorporating of new facts χ_d into a supposedly appropriate model M by the operation of estimation which is deductive and is characterized by the addition $\chi_d + A$.

(b) The Physiology of the Brain:

With two kinds of inference to consider it seems of great significance that research, which under the leadership of Roger Sperry has gathered great momentum in the past 20 years, shows that the human brain behaves[43][44] not as a single entity but as two largely separate but cooperating instruments.

In most people the left half of the cerebral cortex is concerned primarily with language and logical deduction, which plays a major role in estimation, while the right half is concerned primarily with images, patterns and inductive processes, which play a major role in criticism. The two sides of the brain are joined by millions of connections in the corpus callossum, where information exchange takes place. It is hard to escape the conclusion that the iterative inductive-deductive process of discovery is indeed wired into us.

It is well known, that while the left brain plays a conscious and dominant role, one may be quite unaware of the working of the, less assertive, right brain. For example, the apparently instinctive knowledge of what to do and how to do it, enjoyed by an experienced tennis player, comes from the right brain. It is significant that this skill may be temporarily lost if we invite the tennis player to explain how he does it, and thus call his left brain into a dominant and interfering mode.

In this context we see the data analyst's insistence on "letting the data speak to us" by plots and displays as an instinctive understanding of the need to encourage and to stimulate the pattern recognition and model generating capability of the right brain. Also it expresses his concern that we will not allow our pushy deductive left brain to take over too quickly and so perhaps to forcibly produce unwarranted conclusions based on an inadequate model.

While the accomplishment of the right brain in finding patterns in data and residuals is of enormous consequence in scientific discovery, some check is obviously needed on its pattern-seeking ability; for common experience shows that some pattern or other can be seen in almost any set of data or facts. A check that we certainly apply in our everyday life is to consider whether what has occurred is really exceptional in the context of some relevant *reference set* of circumstances. Similarly in statistics diagnostic checks and tests of fit require, at a formal level, frequentist theory significance tests for their justification.

(c) The Mathematics of Bayes Theorem:

It would seem reasonable to require that by a statistical model M we mean a complete probability statement of what is currently supposed to be known a priori (that is, tentatively entertained) about the mode of generation of data $\underset{\sim}{y}$ and of the uncertainty about the parameters $\underset{\sim}{\theta}$ given the assumptions A of the model. At some stage i of an investigation the current model M_i would therefore be defined as

$$p(\underset{\sim}{y},\underset{\sim}{\theta}|A_i) = p(\underset{\sim}{y}|\underset{\sim}{\theta},A_i)p(\underset{\sim}{\theta}|A_i)$$

which can alternatively be factorised

$$p(\underset{\sim}{y},\underset{\sim}{\theta}|A_i) = p(\underset{\sim}{\theta}|\underset{\sim}{y},A_i)p(\underset{\sim}{y}|A_i) .$$

The last factor in the second expression is the predictive distribution. This is the distribution of all possible samples $\underset{\sim}{y}$ which could occur if the model M_i were true.

After the actual data $\underset{\sim}{y}_d$ become available

$$p(\underset{\sim}{y}_d,\underset{\sim}{\theta}|A_i) = p(\theta|\underset{\sim}{y}_d,A_i)p(\underset{\sim}{y}_d|A_i)$$

The first factor on the right is now the posterior distribution of $\underset{\sim}{\theta}$ conditional on the proposition that the actually occurring data $\underset{\sim}{y}_d$ are a realisation from the predictive distribution which results from the assumptions of the theoretical model M_i. If we accept this proposition, all that can be said about $\underset{\sim}{\theta}$ must come from this posterior distribution, and the predictive density is without informational content. However, if, as is always in practice the case, the proposition may be seriously wrong then, correspondingly, residual information may be contained in the predictive density and this can not only indicate inadequacy but even point to its nature. In particular the relevance of the model may be called into question by an unusually small value for the predictive density for the observed sample $\underset{\sim}{y}_d$ as measured for example by

$$\Pr[p(\underset{\sim}{y}|A) < p(\underset{\sim}{y}_d|A)]$$

or by an unusually small value of the predictive density $p\{g(\underset{\sim}{y}_d)|A\}$ of some suitable checking function $g(\underset{\sim}{y}_d)$ as measured by

$$\Pr[p\{g(\underset{\sim}{y})|A\} < p\{g(\underset{\sim}{y}_d)|A\}]$$

Figure 1 illustrates the idea for a single parameter θ and a single *observation* y_d. The particular case illustrated is one where, after the data have become available, it would seem more appropriate to investigate further the adequacy of the model, rather than to proceed with the estimation of θ from its posterior distribution.

There are many conclusions that flow from this approach which are discussed and illustrated elsewhere. The most important in the present context is that the investigational background against which Statistics is applied seems to require that when Bayes' procedure is employed the proposition on which it is conditioned ought to be considered in the light of the data. This can be done by appropriate consideration of the predictive density associated with the data $\underset{\sim}{y}_d$. Such an approach can for example justify and suggest appropriate analyses of residuals, and at a more formal level produces sampling theory significance tests.

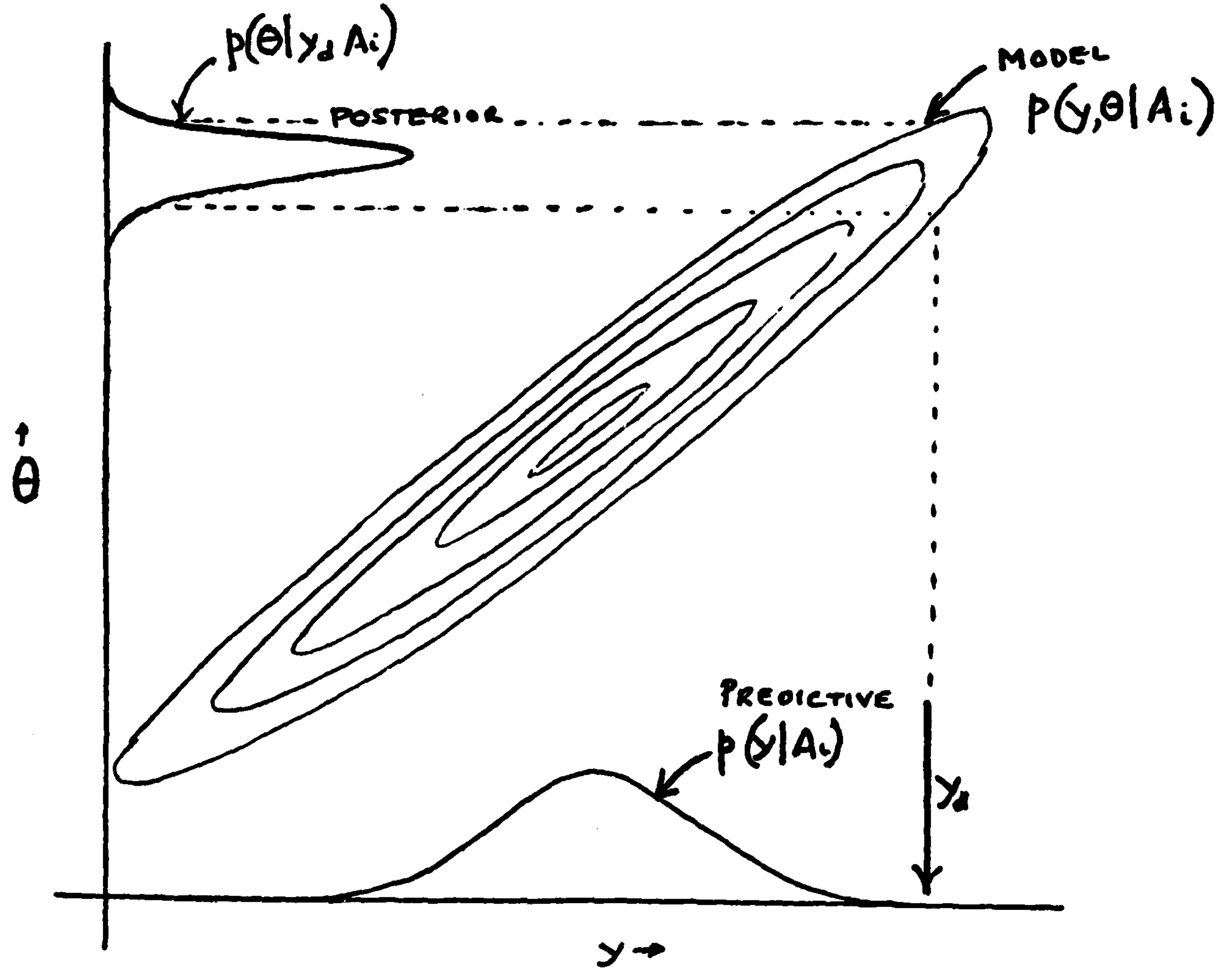


Figure 1. A model with corresponding predictive distribution. The observation y_d produces the posterior distribution shown; however the unusually low predictive density associated with y_d calls the model into question.

5. CONCLUSION

In summary, then, I have tried to show how application and consideration of the scientific context in which Statistics is used can initiate important advances such as: least squares, ratio estimators, correlation, contingency tables, studentization, experimental design, the analysis of variance, randomisation, fractional replication, variance component analysis, bioassay, limits for a ratio, quality control, sampling inspection, non-parametric tests, transformation theory, ARIMA time series models, sequential tests, cumulative sum charts, data analysis plotting techniques, and a resolution of the Bayes - frequentist controversy.

It appears that advances of this kind are frequently made because practical context reveals a novel formulation which eliminates an unnecessarily limiting framework.

REFERENCES

(1) Cochran, W. G. (1978), "Laplace's Ratio Estimator" in Contributions to Survey Sampling and Applied Statistics, Papers in Honor of H. O. Hartley, H. A. David, ed., New York: Academic Press, 3-10.

(2) Galton, F. (1886), "Family Likeness in Stature", Proc. Roy. Soc. of London, 40, 42-63.

(3) Haldane, J. B. S. (1970), "Karl Pearson, 1857-1957" in Studies in the History of Statistics and Probability, E. S. Pearson and M. G. Kendall, eds., Darien, Conn: Hafner, 427-438.

(4) Gosset, W. S. to R. A. Fisher, 15 Dec. 1918, Letter No. 2, in Letters from W. S. Gosset to R. A. Fisher, 1915-1936, (2nd ed.), 1970, privately circulated.

(5) Box, J. F. (1978), Fisher, the Life of a Scientist, New York: John Wiley & Sons, Inc.

(6) Fisher, R. A. (1921), "Studies in Crop Vairation I. An examination of the Yield of Dressed Grain from Broadbalk", Journal of Agricultural Sciences, 11, 109-135.

(7) ____________ (1924), "Studies in Crop Variation III. The Influence of Rainfall on the Yield of Wheat at Rothamstead", Philosophical Transactions of the Royal Society of London, B, 213, 89-142.

(8) ____________ (1926), "The Arrangement of Field Experiements", Journal of the Ministry of Agriculture, 33, 503-513.

(9) Fisher, R. A. and Mackenzie, W. A. (1923), "Studies in Crop Variation II. The Manurial Response of Different Potato Varieties", Journal of Agricultural Science, 13, 311-320.

(10) Wold, H. (1966), "Nonlinear Estimation by Iterative Least Squares Procedures" in Research Papers in Statistics, Festschrift for J. Neyman, F. N. David, ed., New York: John Wiley & Sons, Inc., 411-444.

(11) Yates, F. (1970), Experimental Design: Selected Papers of Frank Yates, Darien, Conn.: Hafner.

(12) Youden, W. J. (1937), "Use of Incomplete Block Replications in Estimating Tobacco-mosaic Virus", Contributions from Boyce Thomson Institute IX, 91-98.

(13) Fisher, R. A. (1935), The Design of Experiments, Edinburgh: Oliver and Boyd.

(14) Finney, D. J. (1945), "Fractional Replication of Factorial Arrangements", Ann. Eugen., 12, 291-301.

(15) Plackett, R. L. and Burman, J. P. (1946), "The Design of Optimum Multifactorial Experiments", Biometrika, 33, 4, 305-325.

(16) Tippett, L. H. C. (1935), "Some Applications of Statistical Methods to the Study of the Variation of Quality of Colton Yarn", Jour. Roy. Stat. Soc., B, I, 27-55.

(17) Gaddum, J. H. (1933), Reports on Biological Standards. III. Methods of Biological Assay Depending on a Quantal Response, Medical Research Council, Special Report Series, No. 183.

(18) Bliss, C. I. (1935), "The Calculation of the Dosage-Mortality Curve", Ann. App. Biol., 22, 134-137.

(19) Finney, D. J. (1952), Statistical Method in Biological Assay, London: Charles Griffin & Co.

(20) Fieller, E. C. (1940), "The Biological Standardisation of Insulin", Jour. Roy. Stat. Soc. Suppl., 7, 1-64.

(21) Daniels, H. E. (1938), "Some Problems of Statistical Interest in Wool Research", Jour. Roy. Stat. Soc., B, V, 89-112.

(22) Pearson, E. S. (1938), In the discussion of Daniels' paper.

(23) Shewhart, W. A. (1931), Economic Control of Quality of Manufactured Product, New York: D. Van Nostrand Co., Inc.

(24) Dodge, H. F. (1969-1970), "Notes on the Evolution of Acceptance Sampling Plans", Journal of Quality Technology, I, April 1969, 77-88; II, July 1969, 155-162; III, October 1969, 225-232; IV, January 1970, 1-8.

(25) Tukey, J. W. (1977), Exploratory Data Analysis, Reading, MA: Addison-Wesley.

(26) Mosteller, F. and Tukey, J. W. (1977), Data Analysis and Regression, Reading, MA: Addison-Wesley.

(27) Wilcoxon, F. (1949), Some Rapid Approximate Statistical Procedures, Stamford, CT: American Cyanamid Company

(28) Bartlett, M. S. (1936), "The Square Root Transformation in Analysis of Variance", Supplement to the Jour. of the Roy. Stat. Soc., 3, 68-78.

(29) Yule, G. U. (1927) "On a Method of Investigating Periodicities in Disturbed Series", Phil. Trans., A226, 207.

(30) Slutsky, E. (1927), "The Summation of Random Causes as the Source of Cyclic Processes" (Russian), Problems of Economic Conditions, 3, 1. English translation in Econometrica, (1937) 5, 105.

(31) Wold, H. (1954), A Study in the Analysis of Stationary Time Series, Uppsala, Sweden: Almquist Wiksell Book Co.

(32) Holt, C. C. (1957), "Forecasting Trends and Seasonals by Exponentially Weighted Moving Averages", O.N.R. Memorandum, No. 52, Carnegie Institute of Technology.

(33) Winters, P. R. (1960), "Forecasting Sales by Exponentially Weighted Moving Averages", Management Sci., 6, 324-342.

(34) Muth, J. F. (1960), "Optimal Properties of Exponentially Weighted Forecasts of Time Series with Permanent and Transitory Components", Jour. Amer. Stat. Assoc., 55, 299-306.

(35) Wallis, W. A. (1980), "The Statistical Research Group 1942-45", Jour. Amer. Stat. Assoc., 75, 320-335.

(36) Wald, A. (1947), Sequential Analysis, New York: Wiley.

(37) Barnard, G. A. (1946), "Sequential Tests in Industrial Statistics", Jour. Roy. Stat. Soc., B, 8, 1-21.

(38) Page, E. S. (1954), "Continuous Inspection Schemes", Biometrika, 41, 100-115.

(39) Barnard, G. A. (1959), "Control Charts and Stochastic Processes", Jour. Roy. Stat. Soc., B, 21, 239-271.

(40) Daniel, C. (1976), Applications of Statistics to Industrial Experimentation, New York: Wiley.

(41) Box, G. E. P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness", Jour. Roy. Stat. Soc., A, 143, 383-430.

(42) Box, G. E. P. (1983) "An Apology for Ecumenism in Statistics", in Scientific Inference, Data Analysis and Robustness, G. E. P. Box, T. Leonard and C. F. Wu, eds., New York: Academic Press.

(43) Blackeslee, T. R. (1980), The Right Brain, Garden City, New York: Anchor Press/Doubleday.

(44) Springer, S. P. and Deutsch, G. (1981), Left Brain, Right Brain, San Francisco: W. H. Freeman.

GEPB/ed

19. KEY WORDS (cont.)

transformation theory, ARIMA time series models, sequential tests, cumulative sum charts, data analysis plotting techniques, Bayesian inference, sampling theory.

20. ABSTRACT (cont.)

data analysis plotting techniques, and a resolution of the Bayes - frequentist controversy.

It appears that advances of this kind are frequently made because practical context reveals a novel formulation which eliminates an unnecessarily limiting framework.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER 2471	2. GOVT ACCESSION NO. AD-A127705	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) THE IMPORTANCE OF PRACTICE IN THE DEVELOPMENT OF STATISTICS		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) George E. P. Box		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709		12. REPORT DATE January 1983
		13. NUMBER OF PAGES 24
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Practice, theory, least squares, ratio estimators, correlation, contingency tables, studentization, experimental design, the analysis of variance, randomisation, fractional replication, variance component analysis, bioassay, limits for a ratio, quality control, sampling inspection, non-parametric tests, (cont.)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The paper shows how application and consideration of the scientific context in which Statistics is used can initiate important advances such as: least squares, ratio estimators, correlation, contingency tables, studentization, experimental design, the analysis of variance, randomisation, fractional replication, variance component analysis, bioassay, limits for a ratio, quality control, sampling inspection, non-parametric tests, transformation theory, ARIMA time series models, sequential tests, cumulative sum charts, (cont.)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

5